

Inference With Contingency Tables (Part 2)

- D. Decision Errors and Power
- E. Significance vs. Effect Size
- F. The Difference Between Testing Proportions vs. Odds Ratios
- G. Confidence Intervals
- H. Larger Tables
- I. Other Models of Hypothesis Testing

D. Decision Errors and Power

In the "decision model" of hypothesis testing (what I referred to previously as the "standard model"), the p-value is used to make a decision between the null and alternative hypotheses, with some appropriate hedging. If the p-value is greater than alpha, the decision is in favor of the null hypothesis. For various reasons, one never accepts the null hypothesis but only fails to reject it, thus acknowledging the fact that no experiment can show that the null is true. If the p-value is less than alpha, the decision is in favor of the alternative hypothesis, although once again the proper conclusion is that evidence in favor of the alternative has been found, and not that the alternative is accepted as true. These are statistical decisions based on probabilities, or likelihoods, and as such we must always remember that our decisions may be in error.

Two kinds of errors can be made, as illustrated in the following table.

	Null hypothesis is true	Alternative hypothesis is true
Null hypothesis retained	correct decision	Type II Error
Null hypothesis rejected	Type I Error	correct decision

Alpha is the Type I Error rate when the null hypothesis is true. That is:

$$\alpha = \text{Pr}(\text{Type I Error} \mid H_0)$$

To put it another way, if alpha is set at the traditional value of .05, then 5% of true null hypotheses will be rejected. That does not mean, as is sometimes stated, that 5% of results in the experimental literature are false. It does mean that a smaller percentage--hopefully, a much smaller percentage--of them are wrong, because the null hypothesis was rejected when it shouldn't have been.

The Type II Error rate is called beta. That is:

$$\beta = \text{Pr}(\text{Type II Error} \mid H_1)$$

In other words, when the alternative hypothesis is true, we will fail to find it so 100*β% of the time. Unlike alpha, which is set by the experimenter at step two of the hypothesis testing process, beta is more difficult to determine.

The ability of a hypothesis test to find a true alternative is called the *power of the test*. Thus, the power of a hypothesis test is 1 - β. We will discuss power more completely at a later time.

E. Significance vs. Effect Size

Power of a test is determined by many factors, one of which is the size of the effect we are looking for. Big effects are easier to detect than small effects. A buffalo is easier to detect than a beetle!

The size of the test statistic and p-value tell us very little about effect size. This is easily demonstrated with a simple coin-tossing demonstration. Suppose we have a coin, and we state as a null hypothesis that the coin is fair. That is, in the long run coin tosses with the coin will yield an equal number of heads and tails.

However, suppose in fact that the coin is not fair but slightly unbalanced to the extent that in every 100 tosses we get one heads too many. Thus, in 100 tosses, the coin should land 51 heads and 49 tails, and in 1000 tosses it should land 510 heads and 490 tails. I think anyone would agree that this is a fairly slight imbalance in the coin--a small effect. In fact, it will be quite hard to detect without a large sample of tosses. The following table shows the expected result of different sample sizes.

no. of tosses	heads	tails	chi-sq	p
100	51	49	0.04	0.841
1000	510	490	0.4	0.527
10000	5100	4900	4	0.046
100000	51000	49000	40	~0

At 100 tosses, the effect isn't apparent by chi square goodness of fit test. Nor is it at 1000 tosses. At 10,000 tosses the effect just sneaks in under the traditional alpha level for significance. At 100,000 tosses. the chi-squared value has grown quite large and the p-value is tiny. In fact, we can make the test statistic as large as we want and the p-value as small as we want by increasing the sample size. The effect size remains the same, however--one additional heads in every 100 tosses.

Statistical significance is not a measure of effect size. We can, likewise, have quite a large effect and yet not find it to be statistically significant if the sample size is too small. Suppose our biased coin lands heads up 80% of the time. That's quite a substantial imbalance in the coin, one that would quickly become noticeable to anyone watching carefully. Yet, if the coin yielded 8 heads in 10 tosses, the chi square goodness of fit test would find $p = .058$, not statistically significant.

Clearly, we need *measures of effect size* in addition to measures of statistical significance. For 2x2 contingency tables there are three commonly used measures of effect size.

- 1) likelihood ratio
- 2) odds ratio
- 3) phi coefficient

Likelihood ratios and odds ratios have already been discussed. The phi coefficient is a simple statistic that combines the chi square statistic with the sample size to give a measure of effect size that is always between 0 and 1. A phi of zero indicates no effect, while a phi of one indicates a perfect effect, i.e., complete agreement between the IV and DV.

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

This statistic is only applicable to 2x2 contingency tables. It does not work with the goodness of fit test or with larger tables, although there is a modification of it that works with larger tables.

Thus, on the Titanic, survival by sex looked like this.

Sex	Survived	
	No	Yes
Male	1364	367
Female	126	344

This is quite a substantial bias in favor of the survival of female passengers, noticeable even from casual inspection of the table. The likelihood of a female passenger surviving was 3.45 times the likelihood of a male passenger doing so. The odds ratio is even more impressive, 10.15 in favor of the female passengers. The chi-squared statistic is 456.87 (and $p \sim 0$). Dividing this by 2201 passengers and taking the square root gives $\phi = .46$. (Note: the phi coefficient is generally calculated from a chi-squared value derived from a test in which the Yates' correction has not been applied.)

If we reduce the frequencies in the table above to 1/100th their current values, the chi squared statistic is much smaller (4.57) and now just significant ($p = .033$), but the effect size measures remain the same. Try it! A big effect is a big effect regardless of how significant it is.

F. The Difference Between Testing Proportions vs. Odds Ratios

Suppose an intervention has been developed for children at risk for the development of attention deficit hyperactivity disorder. We don't know how many of these "at risk" kids will eventually be diagnosed with ADHD, but we think the intervention will save at least some of them from this diagnosis, so we design an experiment with treatment and control groups. The intervention is administered to the treatment group at a young age, and then three years later, after the kids have started school, we check to see how many have been diagnosed with ADHD. Here are some hypothetical results.

group	outcome	
	ADHD	no. ADHD
control	100	100
treatment	90	110

In the control group, 50% of the kids (100 of 200) have ADHD. In the treatment group it's 45% (90 of 200). Thus, the intervention appears to have saved 10 kids in the treatment group from developing ADHD, which is to say 10 of 100, or 1 in 10 kids saved who would otherwise have been diagnosed.

Now suppose the result have turned out this way.

group	outcome	
	ADHD	no. ADHD
control	20	180
treatment	10	190

Ten percent of the kids in the control group (20 of 200) have ADHD, while in the treatment group it is only 5% (10 of 200). The difference between the two groups is still five percentage points. (I.e., the difference in proportions is still .05). However, in the second case, it appears that the intervention saved 10 kids out of 20 from getting ADHD, or 1 in 2 kids saved. Which effect would you say is more important?

The difference in proportions is the same for both groups. It appears that the intervention reduced the proportion of kids getting ADHD by .05, from .5 to .45 in the first case, and from .1 to .05 in the second case. However, I think any reasonable person can see that saving 1 in 2 kids from getting ADHD is a much more impressive result than saving 1 in 10. The difference in proportions does not tell the story. That is, the difference in proportions is not a good measure of effect size.

The sample effect size statistics are shown in the following table.

	LR	OR	phi
case 1	1.11	1.22	.05
case 2	2.00	2.11	.095

In each case, the effect size statistics show that the second effect is larger than the first effect, almost twice as large in fact. It is left as an exercise for the student to determine if in either case the effect is statistically significant. To what extent are the significance tests tests of differences in proportions, and to what extent are they tests of the odds ratios? (This is a trick question.)

G. Confidence Intervals

Another way to make inferences from a sample to a population is by calculating *confidence intervals*.

If a random sample is taken from some population, something is measured about the members of the sample, and a sample statistic is calculated, the sample statistic is said to a *point estimate* of the corresponding population parameter. For example, if we take a random sample of college students, give them all an IQ test, and then calculate a sample mean IQ, that sample mean (the sample statistic) is a point estimate of the population mean IQ (the population parameter). If we take a random sample of American voters, ask them who they favor in the next election, and then calculate the sample proportion who favor candidate XYZ (the sample statistic), this sample proportion is not interesting in itself but is interesting because it is a point estimate of the population proportion in favor of XYZ (the population parameter).

Of course, the point estimate is probably wrong. Since it comes from a random sample, there is random noise, or random error, in the sample statistic. To determine how much random error there might be, we generally calculate another statistic called the sample *standard error* (or estimated standard error). One use for the standard error, as we've already seen, is in a significance test. It often forms the denominator of our test statistic.

All sample statistics have *sampling distributions*. That's just a fancy way of saying, if we take another sample and recalculate the sample statistic, we probably won't get the same result, because of random error. If we take a very large number of random samples and calculate the sample statistic for each one,

we end up with a distribution of sample statistics, say sample means or sample proportions. If we can imagine that we've taken all possible samples, then the resulting distribution (think histogram) of the sample statistic is called the sampling distribution of the sample statistic.

Things get very much easier at this point if we can assume the sampling distribution is a normal distribution, which is not always the case. Oftentimes however, we can assume a normal distribution provided the sample size is large enough. For example, we can assume the sampling distribution of the mean is normal provided we have a sample size of 30 or more, and we can assume the sampling distribution of a proportion is normal provided we have a sample size of 100 or more and the sample proportion is not too close to 0 or 1.

The standard deviation of the sampling distribution is the standard error for the sample statistic .

It sounds like we've got this inference business all wrapped up. All we need is the sampling distribution of the sample statistic, and we know what we need to know. Unfortunately, we don't often have a large number of samples from which to derive a sampling distribution. We have one sample. From that one sample we have to estimate (make our best guess at) the sampling distribution. The math behind doing so can become quite complex and is generally beyond the scope of this course, but a great deal of statistical research has gone into being able to estimate sampling distributions.

From the sampling distribution we can get a second kind of estimate of the population parameter, called an *interval estimate*. One type of interval estimate is called a *confidence interval*. Most commonly calculated is the 95% confidence interval, or 95% CI. If we can assume the sampling distribution is normal, a close estimate to the 95% CI is given by:

$$\text{sample statistic} \pm 2 * \text{se}$$

Take the sample statistic and subtract two times the standard error to get the lower bound of the confidence interval. Take the sample statistic and add two times the standard error to get the upper bound of the confidence interval. (The actual multiplier of the standard error is 1.96, which is the value of z from the unit normal distribution that cuts out the middle 95% of the normal distribution. So if you want to be a stickler for spurious accuracy, multiply by 1.96 instead of 2. I generally let a computer do the arithmetic for me anyway, so consider the method given above to be conceptually correct, although it is not quite technically accurate.)

So a confidence interval is an interval calculated from sample data, and will generally be different for different samples from the same population. It is an interval that is likely to contain the population parameter of interest. The *confidence level* can be set at any value we like (except 100%), but as I mentioned above, it is usually set at 95%. The technical interpretation of this is as follows: given the correctness of the statistical model (i.e., all assumptions correct, etc.), 95% of intervals calculated in this fashion from random samples will contain the population parameter of interest. However, we usually just say that we are 95% confident that the calculated interval contains the true value of the population parameter. (The statement is usually made in terms of confidence and not likelihood or probability. Statisticians frown on probability statements about population parameters.)

It's time for an example, I think! And we'll let R do the math for us. Recall the Doob and Gross study of horn honking as a function of status of frustrator. Motorists were stuck at an intersection behind a new luxury car or an old junker, and the researchers recorded who honked. The results were:

```
> honk.table
              result
frustrator   honk no.honk
  high.status    18     18
  low.status     32     6
```

A chi-square test of independence (without Yates' correction) reveals:

```
> chisq.test(honk.table, correct=F)

Pearson's Chi-squared test

data:  honk.table
X-squared = 9.8732, df = 1, p-value = 0.001677
```

The p-value is quite small, and so we conclude that the researchers have found evidence that honking is related to frustrator status. The proportion of honkers can be found quickly:

```
> prop.table(honk.table, margin=1)
              result
frustrator   honk  no.honk
  high.status 0.5000000 0.5000000
  low.status  0.8421053 0.1578947
```

In the high-status condition, 50% of the drivers honked. In the low-status condition, 84.2% of the drivers honked. The difference in the two proportions is .342. The two-proportion test reveals:

```
> prop.test(honk.table, correct=F)

2-sample test for equality of proportions without continuity
correction

data:  honk.table
X-squared = 9.8732, df = 1, p-value = 0.001677
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.5424007 -0.1418098
sample estimates:
  prop 1    prop 2
0.5000000 0.8421053
```

(In the prop.test, the first column of the contingency table is considered to be successes. Thus, honks are counted as successes. The second column is considered to be failures. The proportion of successes in the second row is subtracted from the proportion of successes in the first row, thus accounting for the negative difference in this case. Since this is a two-sided test, the signs can be ignored.)

The two-proportions test not only gives the same chi-square result but also gives a 95% CI on the difference between the proportions. We now conclude we are 95% confident that the true population

difference in the proportions lies between .142 and .542. This gives us some notion of how accurate the sample difference in proportions might be. Once again, the sample difference is .342, and we are pretty confident that this is within $\pm .2$ of being correct.

Confidence intervals on odds ratios are quite a bit more difficult to calculate, but fortunately we have electronic assistance.

```
> fisher.test(honk.table)

      Fisher's Exact Test for Count Data

data:  honk.table
p-value = 0.002595
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.05239344 0.61835961
sample estimates:
odds ratio
 0.1921131
```

It's easier to talk about these odds ratios if we invert them. (We also need to remember that these are not the sample values but the conditional maximum likelihood odds ratios.) The odds of a low-status car being honked at was $1/.192 = 5.2$ times the odds of a high-status car being honked at. We're none too sure of this value, however, as the 95% CI is $1/.618 = 1.6$ to $1/.0524 = 19.1$, a pretty wide interval of possibilities. Thus, from the given results, we cannot estimate the population odds ratio with much accuracy.

We're pretty confident, however, that it is not 1.0. We're 95% confident that the population value is between 1.6 and 19.1, thus, greater than 1.0. The confidence interval is consistent with our rejection of the null hypothesis from the significance test. This is usually the case. (It is always the case when the CI is based on the normal distribution assumption. It is not always the case when the sampling distribution is not normal and not symmetrical.)

The results of the physicians' study of aspirin and heart attacks was:

```
> aspirin.table
      heart.attack
group   yes    no
placebo 189 10845
aspirin 104 10933
```

The two-proportions test (looking just at the CI) reveals:

```
> prop.test(aspirin.table, correct=F)$conf.int
[1] 0.004687751 0.010724297
attr(,"conf.level")
[1] 0.95
```

Converting to percentages, we can be 95% confident that the true population difference in percentage

of people suffering heart attacks who are on an aspirin regimen versus not is between one-half and one percent. This may not seem like much, and indeed the phi coefficient measure of effect size (calculate it for yourself!) is just .03, a small effect by most standards. However, in this case, even a small effect is an important effect. Roughly speaking, the aspirin saved 85 out of 189 people, or almost 1 in 2 people who would otherwise have had heart attacks, from having one. In terms of the population, that translates into hundreds of thousands or millions of people saved from heart attacks. Small though the effect may be, that's a big deal!

H. Larger Tables

We will now move beyond the special case of 2x2 contingency tables. Most of the basic principles have been laid out, so this should go much more quickly.

For larger tables, the chi-square test of independence is almost always the significance test of choice. (There is a version of the Fisher Exact Test for larger tables, but it is not widely known and rarely used. Proportions can be difficult to derive from larger tables, so the proportions test is usually not used.)

In the R built-in datasets, we find one relating hair color and eye color in a sample of statistic students at the University of Delaware.

```
> HECmale = HairEyeColor[, , 1] # capture all rows, all cols, level 1
> HECfemale = HairEyeColor[, , 2]
> HECmale
      Eye
Hair  Brown Blue Hazel Green
Black   32   11   10    3
Brown   53   50   25   15
Red     10   10    7    7
Blond    3   30    5    8
```

It's pretty tough to see just from eyeballing that table whether these variables are related (for male students in this table). Furthermore, odds ratios and likelihood ratios are out. (How would they be calculated?!) I suggest we start with a graph.

```
> barplot(HECmale, beside=T, legend=T)
```

It appears that brown hair is most common regardless of eye color. It also appears that blond hair is relatively uncommon except in people (men) with blue eyes, and to a somewhat lesser extent men with green eyes. Now that we've characterized the relationship, let's see if it is statistically significant.

```
> chisq.test(HECmale)
```

```
      Pearson's Chi-squared test
```

```
data:  HECmale
```

```
X-squared = 41.2803, df = 9, p-value = 4.447e-06
```

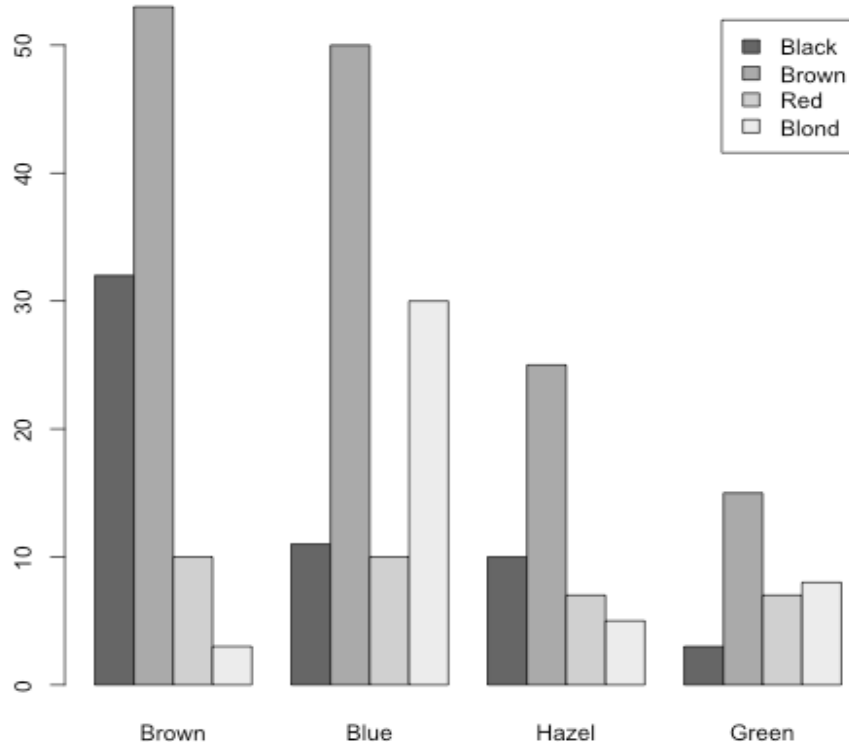
```
Warning message:
```

In `chisq.test(HECmale)` : Chi-squared approximation may be incorrect

Highly ($p < .001$). But what is the deal with that warning message? In this case, R is warning us that some of the EFs have fallen below 5, and that therefore, the chi-squared approximation may be inaccurate. We can check to see which and how many EFs are below 5 fairly easily.

```
> chisq.test(HECmale)$expected
      Eye
Hair   Brown   Blue   Hazel   Green
Black 19.67025 20.27240  9.433692  6.623656
Brown 50.22939 51.76703 24.089606 16.913978
Red   11.94265 12.30824  5.727599  4.021505
Blond 16.15771 16.65233  7.749104  5.440860
```

Only 1 of 16, and only by a little bit. I'm not going to worry too much about it, especially since the p-value is not marginal (it is waaaaaaay below .05).



Effect size measures for contingency tables larger than 2x2 are also problematic. The one most commonly used and most widely known is Cramer's V, which is a modification of the phi coefficient, and so is sometimes also called Cramer's phi. It is always a value between 0 (no effect) and 1.

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

In this calculation, k is the smaller of the number of rows or columns in the table.

```
> sqrt(41.28 / (3 * sum(HECmale)))  
[1] 0.2220788
```

Analysis of the HECfemale table is left as an exercise.

I. Other Models of Hypothesis Testing

The decision model of hypothesis testing, in which the test is set up to allow us to decide between a null and alternative hypothesis, is called the *Neyman-Pearson model* of hypothesis testing. Although it is the "standard model" taught in almost all introductory statistics books these days, it has taken a beating lately in the statistical literature, and is not universally accepted. There are alternatives, of which I will discuss one, called Fisher's (or the Fisherian) model.

One problem with the decision model is that p-values are not very reliable. To suggest that we can decide one way when $p = .04$ and another when $p = .06$, and do so with any credibility, strikes some people (myself included) as far fetched. So why force a decision on ourselves? Why not interpret the p-value as "strength of evidence against a null hypothesis." This is Fisher's model in a nutshell.

In Fisher's model, the null hypothesis (H_0) is stated, the data are collected, and a p-value is obtained from the data. If the p-value is low enough, then the null hypothesis may be rejected. It was Fisher who suggested the customary criterion of $p < .05$ for rejection of the null, but he also at times suggested that $p < .01$ be used, and at other times other values.

At first blush, the difference between the two systems may not seem all that great, or to be rather "philosophical." Perhaps. But it has led to a great deal of debate over the meaning of p-values and over conventions for reporting statistical tests on experimental results. I don't intend to rehash that here. I'll merely repeat: p-values obtained from experimental results are simply not very reliable. It strikes me as being silly to make black-and-white, in some cases even life-and-death, decisions based on a statistic that cannot be replicated or that may change from one side of a cutoff to the other based on very trivial differences in the data. It makes more sense to see "the p-value as a sample-based measure of evidence against the null hypothesis" (Zieffler, et al., 2011, p.178), and to insist that the truth lies in replication and not in a single p-value. Important results should be replicated!

The following table offers qualitative descriptions of how p-values might be interpreted as strength of evidence against a null hypothesis.

p-value	strength of evidence
.1 to .05	weak or marginal
.05 to .01	moderate
.01 to .001	strong
<.001	very strong or overwhelming

(This table is a modification of the one in Zieffler, et al., 2011, p.179.)